The Role of Timing Cues in Speech Perception

By

Jeremy André Lacocque

HONORS THESIS
for the Degree of
Bachelor of Science in Psychology
College of Liberal Arts and Sciences
University of Illinois at Urbana-Champaign

April 2011

Acknowledgments

I would like to thank my advisor, Robert E. Wickesberg, for his unconditional, enthusiastic support since the first day, when I knew next to nothing about speech perception. I would also like to thank Justin Rhodes and Jennifer Fayard for their ongoing support throughout the processes of growing as a researcher, writer and presenter. Finally, I would like to thank Hanna Stevens, who made this project possible by collecting the data used in this research.

TABLE OF CONTENTS

Abstract

Theories of speech perception posit that recognizing key frequencies in speech sounds is the primary way we recognize speech. These theories, however, do not explain how we understand speech with shifted, key-frequency information such as whispered speech. In an effort to determine characteristics of speech vital to speech recognition beyond just frequency, timing patterns were examined in auditory nerve fibers' action potentials throughout the presentation of speech sounds. Since chinchillas have similar peripheral auditory systems as do humans, they were used in the study. Eighteen chinchillas were anesthetized and their auditory nerves exposed to record their activity. Alveolar stop consonants speech sounds /pap/, /paep/, /bab/, and /baeb/ were presented to chinchillas' peripheral auditory systems by a female speaker in two ways: normally spoken, and whispered, each at three different intensities. Despite the fact that whispered speech contains different frequency information than does normally spoken speech, the experiment is expected to reveal that each consonant evokes a similar timing pattern in the auditory nerve fibers across method of speech production, but distinct between consonants, regardless of whether the speech sound was whispered or normally spoken.

Timing patterns in the auditory nerve fibers' responses were gathered and analyzed by pooling the responses from each auditory nerve fiber from each chinchilla to create an ensemble-average, time histogram. For each speech sound, a specific number and pattern of peaks in action potentials at certain points in time appeared in each histogram, sometimes independent of method of speech production, whispered or normally spoken. Certain histograms revealed similar timing patterns encoded in the chinchillas peripheral auditory system for each pair of speech sounds, showing that timing information is consistent between different forms of speech production and therefore might be an important cue in how we process speech.

Introduction

For more than sixty years psycholinguists have developed theories as to how normally voiced speech is perceived, but none explains other forms, such as speech perceived by cochlear implants, sung speech or whispered speech. The existing theories claim normally voiced speech perception relies primarily on frequency information within the speech sounds. In speech perceived by a cochlear implant, or speech that is shouted or whispered, the frequency information is different and the main theories of speech perception relying on consistent frequency information therefore do not address how speech is perceived in these situations. How exactly humans perceive whispered speech without typical frequency information is unaccounted for in the main speech theories and is what this study intends to shed light on. Specifically, whether or not timing information is the same in normal and whispered speech will be examined in an effort to develop a new speech theory, less reliant on frequency information, accounting for every form of speech.

The production of normally voiced speech starts when specialized movements of vocal organs vibrate air and generate sound waves. Normally, when someone exhales, the stream of air is almost inaudible. However, when the air is vibrated by vocal cords, it becomes auditable (Denes & Elliot, 1993). The sound the vibration makes can be changed by using articulators like the palate, tongue, lips and jaw. When a sound is shaped by articulators throughout the vocal tract in certain ways, they produce speech sounds. Whispered speech, on the other hand, relies on the pairing of exhalation with careful articulation of the mouth, lips and tongue while the air exits the mouth, which, again, produces speech sounds. Speech sounds, regardless of how they are produced, eventually produce syllables, words and finally sentences. Several problems arise when modeling an auditory system capable of understanding these speech sounds, however.

Individual components of speech sounds, called phonemes, can often overlap, making it the job of the auditory system to tease them apart so the proper speech sound can be perceived. This overlapping is called coarticulation, and how exactly the auditory system deals with this has been difficult to explain, much less when there is background noise. How the problem of decoding speech sounds and coarticulation is solved in speech perception, and how speech sounds are perceived by the listener and processed into words has created much debate, and is what the theories of speech perception attempt to explain.

The three main theories since the 1950's have explained many aspects of speech perception, but none has been exempt from heavy counterevidence. In 1957, the first theory was developed, called the Motor Theory by Alvin Liberman, which had been left unchallenged for decades and stands as the main theory of speech perception, despite evidence against it. The theory relies on intricate variations in certain frequencies during speech production, which has since been shown to be unnecessary as summarized by Diehl et al. 2004 in a review paper. The second of the three prominent theories, called the Direct Realist Theory, relies on the many of the same principles as the Motor Theory and has likewise been discredited. A third, alternate theory was established twenty years after these two theories took into account the evidence against the first two theories but has failed to provide any new ideas about how speech is perceived. Since the General Auditory Theory does not propose many new concepts, but rather focuses on opposing the previous theories, it is viewed more of as an approach than a theory (Diehl et al., 2004). While there are other theories of speech perception, most fail to provide solid evidence, make unique claims or are able to explain things like noise-vocoded speech and will therefore not be discussed. The gap in knowledge which still needs to be filled is therefore what cues are used to perceive speech given frequency information is not as relevant as

previously thought. By studying speech with little or shifted frequency information, like whispered speech, perhaps new theories can be developed, which is what this study aims to do.

The Motor Theory was the first theory published, and came out in the 1950's. Alvin Liberman and his colleagues were the first to articulate a coherent theory, which is perhaps why it has been so prominent. From 1951 to 1956, Liberman and his colleagues, Dranklin Cooper, Pierre Delattre and others, worked on building the foundation for the theory, including finding that phonemes and speech sounds are closely related to articulatory events and some acoustic information, like variation in frequency (Liberman et al., 1952). In other words, the intention of the gestures used (movement of mouth, tongue, lips) to produce speech are perceived and the listener uses those cues to decode speech sounds, relating those intentions of producing speech to that of his or her own. So, if a speaker produces the sound /da/, the listener recognizes the movement of the lips, mouth and tongue that the speaker must make to produce that sound, and uses that information to understand the speech. The actual acoustic and auditory properties of speech were postulated to be less relevant in perceiving speech, although still important (Liberman, 1957). A major strength of The Motor Theory is this proposal of the listener's internal articulation, which helps compensate for coarticulation and other context-dependent variations in speech as well (Denes & Elliot, 1993).

A few years after Liberman's initial proposal of the Motor Theory, he and his colleagues theorized that, in addition to articulatory cues, frequencies were important cues in speech recognition. He concluded this when he mapped prominent frequencies present in speech called "formants," which was modeled after work from Peterson and Barney, 1952, who established the idea of formants. In spectrograms, essentially visual representations of frequencies over time, these formants are visible as horizontal lines. These lines in the spectrogram indicate a

frequency, more prominent than others, which stay constant over time. This straight, horizontal line is often preceded by a little tail representing a minute shift in frequency, called a "formant transition." This small change in frequency of about 500 Hz, supposedly allows the listener to distinguish between consonants, while the formant itself allows distinction between vowels (Liberman et al., 1967). The listener would first perceive the vowel using the formant, then the preceding formant transition, to construct a consonant-vowel pair leading to a syllable, and eventually a word (Liberman & Mattingly, 1985).

The ability to perceive speech this way, according to Liberman, is one unique to humans. This claim would make sense from Liberman's point of view since perceiving speech requires the listener to understand how to produce the speech the listener hears, something a dog could not do for instance with a different vocal tract than humans have. The ability to perceive speech, however, is not unique to humans as parrots can repeat speech, and pets can understand commands and can therefore indeed perceive speech to some extent. If perceiving speech were unique to humans, dogs could not possibly understand human speech and follow complex commands as they do. The Motor Theory therefore implies that every animal has a different method of perceiving speech, as every animal has a different physical way to product speech, if it is able to at all.

The Direct Realist Theory attempts to take into account that non-humans can perceive speech and was first published in the 1980's by Carol Fowler, a colleague of Liberman. The theory resembles the Motor Theory in many ways, but it differs at a few points, which are summarized by Diehl et al. in a 2004 review paper. For instance, the theory asserts that the physical production of speech sounds are perceived, not necessarily the preceding neuromuscular commands and intentions like postulated in the Motor Theory (Diehl et al., 2004). It also

addresses Liberman's claim that the speech-perception process is unique to humans, and claims the contrary. Fowler then implies the way we perceive speech is most likely not unique to humans, which is where Fowler's Direct Realist Theory differs most starkly with the Motor Theory. The review paper explains key points of the theory by examining the theory's name, "Direct Realist Theory." "Direct," it says, most chiefly refers to the fact that the acoustic signal of the speech is rich enough to supply adequate information about the gestures used to produce the speech, and no inferring must really take place. "Realist" suggests physical, real gestures are perceived and used to decode speech sounds (Diehl et al., 2004).

Both the Motor Theory and the Direct Realist Theory rely on formants and formant transitions, and therefore spectral (frequency) information to decode speech. Again, formants are defined by specific, unchanging frequencies and formant transitions by minute changes in frequency of about 500 Hz. Speech, however, can still be understood without this information, providing evidence against the two theories. In 1995 Robert Shannon and colleagues were one of the first to show that speech could still be understood with limited frequency information, and preserved temporal (timing) information (Shannon et al., 1995). The study took recordings of normally spoken consonants, vowels and sentences and split them up into one, two, three or four frequency bands. Each band was made into "white noise," which replaced any existing frequencies with every frequency in a specific range, depending on the band. Each band consisted of a group of frequencies whose amplitude (volume) was independently modulated from another, ridding the signal of most spectral cues. With as little as three bands, or groups of frequencies, roughly 80 percent of subjects were able to distinguish between consonants, and between sentences. With as little as three bands, about 90 percent of subjects could distinguish between vowels (Shannon et al., 1995; Loebach & Wickesberg, 2006). With only three bands,

detecting the minute fluctuations in frequency found in formant transitions and the frequencies of formants would be likely impossible. Formant transitions occupy about 500 Hz range, while each band occupied a larger range, about 700 to 2,000 Hz, depending on how many bands were present (Shannon et al., 1995). So, 500 Hz formant transitions would not be present within a band, thus discrediting the idea that formants and formant transitions are primarily how speech is decoded, as stated in the Motor Theory and Direct Realist Theory.

A later study by Loebach and Wickesberg performed a similar task as Shannon et al., but specifically studied ensemble responses in the auditory nerve of the chinchilla. Chinchillas are often used in auditory studies due to their auditory system's similarities to that of a human's (Kuhl & Miller, 1975). For many animals, the key difference between its hearing and a human's, is the frequencies it can hear. The chinchilla is often used because its auditory system is sensitive to almost the same frequencies as humans.

Just as in Shannon et al. 1995, Loebach and Wickesberg split speech into different bands of white noise, each independently amplitude-modulated to study the responses of the auditory nerve of the chinchilla (Loebach & Wickesberg, 2006). Ensemble responses revealing temporal patterns of the auditory nerve were computed by averaging values of auditory-nerve-fiber responses. An auditory nerve fiber is one of about 30,000 neurons within the auditory nerve, each responding optimally to a certain frequency (its characteristic frequency) at a certain intensity (its threshold). This averaging was necessary because analyzing responses of individual nerve fibers would be almost meaningless because as frequencies shift, as they do in whispered speech, different auditory nerve fibers would be activated. So, in order to attain information about temporal cues across frequency, it was necessary to look at every nerve fiber's response together. Because the spectral information was degraded by using noise-vocoded speech,

averaging responses across auditory nerve fibers, each responsible for a certain frequency range, allowed temporal information to be carefully examined (Loebach & Wickesberg, 2006). It was found that temporal information in ensemble responses for 3- and 4-band noise-vocoded speech was nearly identical to that of natural speech, indicating temporal information is largely independent of spectral information. Shannon et al., showed that this is enough to allow subjects to perceive most consonants, vowels and sentences (1995). The study concluded that the temporal patterns found from the auditory nerve of a chinchilla were enough to provide the cues necessary to understand noise-vocoded stop consonants, sounds like /p/,/t/,/k/,/g/ and /q/ in English.

The fact that only a few bands are needed to understand most speech is illustrated by how cochlear implants work. The surgically implanted device replaces the job of the cochlea, and transduces air vibrations directly into electrical signals, which then get passed onto respective auditory nerve fibers (Simmons, 1966). A strip of electrodes gets placed into the cochlea, each electrode responsible for stimulating a certain set of auditory nerve fibers. French-Algerian surgeons André Djourno and Charles Eyriès were the first to stimulate auditory nerve fibers using electrodes in the 1950's. In 1957, the scientists developed the first cochlear implant, which had one electrode (Simmons, 1966). While one band was often not enough to understand speech, it helped provide temporal cues to pair with lip reading.

The effect of this electrode is similar to what Shannon, Stevens and Wickesberg did, since cochlear implant relies on a certain number of bands of frequencies to send acoustic information to the hearing-impaired person's brain (Stevens & Wickesberg, 1999; Shannon et al., 1995). So, the number of electrodes the implant had is equivalent to the number of bands, or groups of frequencies in whose amplitude were independently modulated. As in Shannon's

paper, at least three or four bands were needed to begin to understand speech sounds (Shannon et al., 1995). The fact that speech could be understood with only this many electrodes in a cochlear implant, and enhanced with even only one electrode, emphasizes that frequency cues cannot be entirely relied on when decoding speech sounds. Within a couple decades, cochlear implants had up to 24 electrodes, making speech easily understandable and music enjoyable. In addition to this technology, the implant also had a processor, which emphasized certain frequencies to help the listener decode speech more accurately. Frequencies such as those in formants were amplified. While it has been shown formants and formant transitions are not solely sufficient to perceive speech, they certainly can work with other cues to enhance the accuracy of speech perception, which is why they were implemented in cochlear implants. Research about timing cues in speech perception can hopefully lead to better programming of processors for cochlear implants, so more vital aspects of speech can be highlighted, and speech recognition enhanced.

Taking into account the evidence that frequency information is not entirely responsible for speech perception, The General Auditory Theory was formed. The theory is mainly defined by its opposition to the Motor Theory and Direct Realist Theory and is therefore classified as more of an approach than a theory (Diehl et al., 2004). Because of findings like those of Shannon's, speech scientists have explored alternatives to the Motor Theory and the Direct Realist Theory, but have not yet developed an alternative theory, but rather an approach, called The General Auditory Theory. It is considered as more of an "approach" than a theory because it does not offer many new ideas as to how speech is perceived, but rather, opposed many of the previous theories (Diehl et al., 2004). It assumes the speech sounds are perceived the same way other acoustic events are in the environment, and has nothing to do with gestures, or how speech sounds are physically produced (Diehl et al., 2004). The General Auditory Theory has also

asserted that visual information is incorporated with acoustic information to better understand speech. This claim is supported by the McGurk effect, which illustrates that if there is a discrepancy in visual and acoustic information, the auditory system can be tricked, implying that visual information can be influential, if not dominating, in perceiving speech. This aspect of the General Auditory Theory, however, is somewhat discredited in that speech can be understood while on the phone, or with eyes closed, when no visual information is present. It can be argued, however, that visual information is not essential, but rather just an important helping aid.

Another cue said to aid in speech perception is voicing, and voice-onset time, first discussed by Leigh Lisker in 1978, who also worked in the same lab as Liberman and Fowler. Voice-onset time is defined as the time between when the stop consonant (like /p/ and /b/) is released and when voicing begins. According to Lisker in 1978, this measure is the single most adequate physical cue for differentiating between stop consonants in the presence of other features. The study indeed confirmed that alone, voice-onset time is an insufficient cue to distinguish between stop consonants. It is, however, helpful when paired with other cues (Lisker, 1978).

Speech without typical spectral cues and voicing (vibrating of vocal cords) can be understood almost as accurately as normal speech, evidence that, when compared to frequency, temporal cues again probably play the larger role in speech perception. An example of such speech is whispered speech. Whispered speech, which among other things lacks the same harmonic cues as normally voiced speech and the frequency information differ greatly (Schwartz, 1970; Repp & Lin, 1989). This provides evidence against the Motor Theory and Direct Realist Theory, which rely on consistent frequency information. Since the spectral cues in whispered speech differ from that of normal speech, and formant transitions are different, the

Motor Theory and Direct Realist Theory are easily contested when explaining how one can perceive whispered speech. If, however, the differences in frequency between normal and whispered speech could not affect the cues most important for perception, it is possible frequency information could be used to perceive speech (Stevens & Blumstien, 1981). In other words, although the frequency information may be different, there may be mechanisms in the auditory system, compensating for these differences allowing whispered speech sounds to be perceived the same way as normal speech (Tartter, 1989; Dannenbring, 1980).

Another aspect of whispered speech is the absence of voicing. Whispered speech by definition lacks use of vocal cords and the exhaled air is therefore much softer with different auditory properties. Vivien Tartter pointed out in 1989 that when certain consonants are whispered, there are no longer voicing cues and therefore voice-onset time cues. Subjects, however, can still somehow successfully differentiate between them. She then concludes that voicing cues must somehow be present in other forms in whispered speech (Tartter, 1989). Whether or not this is the case, there are clearly other vital cues, present even in whispered speech, which could have been used to differentiate between consonants. Within whispered speech, it is possible that vowel-onset time could be used to differentiate between stop consonants. Instead of depending on the onset of voicing, which whispered speech lacks, the beginning of the vowel would be used as the landmark instead, assuming the time between the consonant and vowel is identifiable. Hanna Stevens and Robert Wickesberg mention temporal cues that are present that may help distinguish between stop-consonants, especially when voice-onset time cues are unavailable (Stevens & Wickesberg, 1999). By examining ensemble responses across frequencies and auditory nerve fibers, temporal cues were found to be virtually the same across whispered and normal speech but unique between consonants, providing a

sufficient-enough cue to distinguish between consonants. Cochlear nucleus neurons also showed a similar temporal pattern (Clarey et al., 2004). This study hopes to show the same timing-pattern consistency across speech production method, but with bilabial stop consonants /p/ and /b/.

Whispered speech and normally voiced speech are often analyzed in these studies with spectrograms, or visual representations of sound. The human brain, however, analyzes sound and processes it differently. The process begins when the sound wave reaches the listener's outer ear or pinna. The shallow, funnel-shaped structure guides the sound wave into the ear canal, where it can eventually reach other important structures, like the ear drum. The pinna also helps the listener determine the direction of the source of sound, as its unique shape and ridges cause sounds to be changed in different ways before entering the ear. The brain can then use these differences in sound caused by the shape of the ear to provide information about the source of the sound. The ridges, or "notches," on the pinna also aid in providing elevation cues, certain high-frequency cues than allow the listener to determine how high or low the source of the sound is. Binaural hearing, or hearing with two ears as opposed to one, also allows the listener to locate the sound. If sound reaches one ear before the other, the brain understands it is most likely coming from that direction.

The sound wave then reaches the tympanic membrane, or eardrum, which is the barrier between the ear canal and middle ear. This membrane vibrates, oscillating back and forth, depending on the frequency and amplitude of the sound. Air is not the only medium sound can move through, however. Fluid also conducts sound well, although it is harder to move than air. Because of this, in order for the fluid in the next part of the ear, the cochlea, to move, the vibrations from the eardrum must be amplified. This happens with the help of three bones, the malleus, incus and stapes, some of the smallest bones in the human body. The bones amplify the

vibrations by about 20 fold using lever-like properties, which then allow the fluid in the cochlea to vibrate. The word "cochlea" is Latin for "snail" because of its snail-like shape, and "pinna" Latin for "feather," describing its shape. The curled-up tube comprising the cochlea has two chambers through which the vibrating fluid travels. This vibrating fluid then vibrates the basilar membrane, a thin membrane separating the two chambers of fluid. This vibrating membrane then leads to the transduction of the mechanical energy of sound into neural impulses when it comes into contact with the Organ of Corti.

Different parts of the basilar membrane have different thicknesses and rigidities, causing different parts of the membrane to vibrate at different frequencies. The rate and location of these vibrations allow corresponding parts of the Organ of Corti to activate as well (Corti, 1851). This structure, covered in thousands of small hair cells, lies on top of the basilar membrane and is what changes the mechanical energy of sound into the neural impulses in the auditory nerve. When the basilar membrane vibrates at a certain spot under it, it moves the small hair cells on the Organ of Corti, each which corresponds to certain auditory nerve fibers. Depending on which auditory nerve fiber is activated and the rate at which it repeatedly activated, the brain can tell which frequencies are present in the original sound. With low frequencies, the auditory nerve fibers fire at the same rate as the frequency. This explains the basis of the Volley Theory. However, for higher frequencies, the firing rate cannot keep up with the frequency of the sound, so the location on the basilar membrane, which vibrates the most, helps the brain determine what frequency is being heard. The brain can then differentiate between one frequency and another based on where it causes the basilar membrane to vibrate the most.

Several theories of speech perception rely heavily on this process, as they claim that minor frequency fluctuations and presentations help us decode speech sounds. Frequency

information is not the only acoustic information our auditory system provides our brain with, however. Temporal information, information about when certain frequencies occur, is also thought to be important (Shannon et al., 1995).

Cells, which reside in the ventral cochlear nucleus, where many of the auditory nerve fibers come together, are thought to be responsible for preserving temporal information. Without these cells, different frequencies would be heard at different times. These cells are called "Octopus cells" and take the information from different auditory nerve fibers and integrate and preserve their timing information, so as to prevent temporal separation. Although not discovered until recently, this temporal information might play a large, and even more important role in speech perception than frequency alone and is something the body takes care to preserve with these cells (Stevens & Wickesberg, 1999).

The whispered alveolar stop consonants /t/ and /d/ were specifically examined in Stevens' and Wickesberg's study when they compared /t/ and /d/ in whispered and normal speech and found similar timing patterns (1999). Bilabial stop consonants /p/ and /b/ have also been examined by Clarey et al., but only during normal speech (2004). The goal of this study is to test the hypothesis that the ensemble responses from the auditory nerve will elicit the same timing patterns for whispered speech and normal speech for bilabial stop consonants /p/ and /b/.

Methods

The following methods were adapted from Stevens and Wickesberg (1999) and Loebach and Wickesberg (2006). The procedures have been approved by the Institutional Animal Care and Use Committee of the University of Illinois at Urbana-Champaign.

Eighteen Chinchillas (Chinchilla laniger) were anesthetized with ketamine HCl (40 mg/kg), xylazine (2 mg/kg), and acepromazine maleate (4 mg/kg) to make sure they were unconscious and felt no pain during the invasive procedures of the experiment. Throughout the experiment on the chinchillas, the animals received additional doses of the drugs to maintain their unconscious state. A small hole was made in the chinchilla's airway before surgery and the animal's head was fixed in a custom-designed nose clamp to maintain the animal's airway. For the duration of the experiment, the internal temperature was maintained at its normal body temperature of 36 degrees Celsius with a feedback-controlled heating pad. The entire experimental setup was located in a shielded, sound-proof chamber to isolate the sound used in the experiment from sounds and vibrations from the outside environment. The animal's right ear-canal opening was exposed by removal of the bony wall around it, to allow a direct approach to the ear drum, at the other end of the ear canal. An earpiece, each with two small speakers inside used to deliver the speech sounds were fitted with a foam tip and inserted in the open canal to within a few millimeters of the ear drum and then sealed with petroleum jelly to ensure maximal sound quality. To ensure accuracy of intensity level, these earphones were calibrated using a small microphone. The middle ear cavity typically has a tube, called the Eustachian tube, to equalize the pressure on either side of the ear drum. Even though only healthy animals were selected, this tube can often get clogged during mild sicknesses or infections, so an artificial tube was inserted through the top of the skull and the passage of air maintained by an 18-gauge needle. Without equalized pressure on either side of the ear drum, sound would not get properly propagated from the ear canal to the inner ear, where the sound gets amplified and eventually transduced in the cochlea.

The auditory nerve, connecting the inner ear to the brain, was approached from behind by removing a portion of the overlying brain tissue to leave a direct view of the dorsal cochlear nucleus. This gave clear access to the auditory nerve, and therefore the auditory nerve fibers, which we probed during the presentation of speech sounds to gather timing information. The auditory nerve fibers are the neurons within the auditory nerve that begin at the hair cells in the cochlea and travel to the brain. It is these auditory nerve fibers we examined and recorded responding at certain times in response to certain speech sounds. When a sound is transduced into electrical signals in the cochlea, those neural signals tend to occur in different patterns and at different times throughout the speech sound. Each "neural signal" is an action potential. The number of action potentials within a half-millisecond time period were measured for the duration of the speech sound presentation.

To record these responses to try and analyze patterns in timing, glass microelectrodes were placed against the outside of the nerve fiber and detected small voltage changes. Since glass does not conduct electricity well, the electrodes were filled with electrolytes potassium and chloride, which left the probe with an impedance of approximately 20 MT. A microelectrode was entered into the chinchilla aimed toward the visible portion of the auditory nerve or the most lateral part of the exposed portion of the ventral cochlear nucleus. The area of recording was from fibers immediately after they exit the internal auditory opening of the lateral bony wall that separates the brain stem from organs of the inner ear and immediately before they enter the ventral portion of the cochlear nucleus. As a method of distinguishing between neurons in the auditory nerve and neurons in the cochlear nucleus, clicks were presented to the chinchilla, and neurons responding with a latency to the clicks of less than 4 ms were classified as auditory

nerve fibers, and those with a latency of more than 4 ms classified as cochlear nucleus nerve fibers.

Before recording, the electrode was surrounded and stabilized with type I agarose (4% solution). Because the voltages were gathered from outside the nerve cell where they were weaker, instead of ideally on the inside where distinct, strong differences in voltage would be detectable, the recorded waveforms had to be amplified. Once they were amplified, a machine was calibrated to the waveform of an action potential and used to filter out any irrelevant "electrical noise" that were not action potentials. The speech sounds were produced by a computer, where normally spoken or whispered speech sounds /bab/, /baeb/, /pap/ and /paep/ were previously recorded and three intensity levels, 60-, 70- and 90 dB pe SPL. This same program collected the data on when these auditory nerve fibers had action potentials. With each auditory nerve fiber, a plot of how that neuron responded to a set of frequencies, was examined. The graph, called a "tuning curve" would reveal a peak, a frequency at which that neuron responded most optimally, called the characteristic frequency. Because each neuron has a certain number of action potentials even in the absence of stimuli, the "spontaneous rate," or rate at which a neuron fires in the absence of stimuli, a value for that was recorded as well. The threshold for the neuron to be most active was also measured and recorded. The collection, presentation, and storage of these data from the experiment were available off-line with a PC version of the Response Analysis Package from the Department of Neurophysiology at the University of Wisconsin-Madison.

Eight different speech stimuli were created by recording a female voice speaking and whispering eight different speech sounds, each beginning with an alveolar stop consonant, /p/ or /b/. The specific consonant-vowel syllables used were /baeb/, /bab/, /paep/ and /pap/ in both

normally spoken, and whispered forms. The speaker was recorded with a microphone held five inches from her lips. The syllables were sampled using a Labtec dynamic microphone and a Diamond sound card in a 75 MHz Pentium computer. The sampling rate of the recorded sounds was 44.1 kHz and the recordings were stored as waveform files. The same stimulus was recorded at least eight times from the speaker in each manner of voicing, the normal speaking voice and the whispered voice. The samples were analyzed for clarity, distortion and variations in the waveform. The most distinct version was chosen to represent the stimulus type. The final stimuli consisted of a 2-ms silence followed by the entire waveform of the speech sound. In the presentations, the stimuli were repeated 50 times in immediate succession with a silence interval equal to the length of the sound itself. While all auditory nerve fibers or the hair cells may not have recovered even halfway, the repetition rate of the speech sounds was much slower than the natural presentation of speech. Listening to the stimuli**,** we determined that they are easily distinguished even with no recovery time between presentations.

Both whispered and normally voiced sounds were analyzed to determine their intensities. The intensity of presentation was determined by the peak intensity of the speech sound. Stimulus intensity was calculated from the peak amplitude of the sound measured in the ear canal with the probe microphone using the amplitude of a certain, pure tone as the reference. Since the intensity is based on the equivalent of the peak, it will be referred to as the peak equivalent SPL or "pe SPL." The intensity of the unchanged speech was 100 dB pe SPL. These are the intensity values used in this paper. The intensity of a stimulus was also calculated as the average power during the long part of the vowel sound.

After initial characterization of an isolated auditory nerve fiber, the normally voiced speech sounds from the female speaker was then presented 50 times to the chinchilla via the two,

small speakers inside its ear at levels of 60, 70 and 90 dB pe. 60 dB is just a little louder than the typical volume of normally spoken speech, and 90 about as loud as a concert.

The timing of action potentials produced by an auditory nerve fiber in response to one of eight speech sounds was recorded, when either whispered or normally spoken. It is this timing data we hope present similar patterns for both whispered and normally voiced speech. The times of the action potentials were used to compute "per-stimulus time" graphs, which graphed how many times an auditory nerve fiber had an action potential over time in response to a speech sound. These graphs will be called "PST histograms" throughout this paper.

The data from RAP, (the characteristic frequency, threshold and spontaneous rate of each auditory nerve fiber) were then matched with their respective timing data collected from probing each nerve fiber. The timing data was represented by a number of action potentials per half millisecond time period. Because, as mentioned before, each neuron has its own spontaneous rate, or tendency to have an action potential even in the absence of input, the spontaneous rate was subtracted from the action potential data so only the number of action potentials directly in response to the speech sound were theoretically present. Without this, the neurons with high spontaneous rates would appear to be much more responsive due to the speech sound than they really were, which would skew the data. These data were then organized by intensity, 60, 70 or 90 dB pe SPL. Once each auditory nerve fiber was matched with its properties and action potential data, the average number of action potentials for a certain half-millisecond period were then calculated. Characteristic frequencies falling outside of the range of 250 – 7000 Hz were omitted, leaving us with just the "ensemble response" described in the introduction of this paper. These omissions were based on the articulation index (AI) described by French and Steinberg (1947). The AI posits that the most important information for the perception of speech is

contained between the frequencies of 250 and 7000 Hz, and therefore named that range the articulation index. We, therefore, discarded any information outside this range, as it is assumed it is irrelevant. The articulation index frequency range was divided into 20 bands, each which were assumed to contribute 5 percent of the relevant speech sound information. This process was done to create an ensemble average that despite sampling variations would approximate an average in which each frequency band contained an equal numbers of auditory nerve fibers with the same distribution of spontaneous rates. The data was then organized by characteristic frequency distribution and plotted on a logarithmic scale. Because the distribution was not evenly distributed on the logarithmic scale, the data was normalized to fit the scale. Without an even distribution of spontaneous rates or characteristic frequencies, the PST histograms would not have been comparable to one another. This process essentially standardized, or normalized, the data such that one histogram could be compared to another qualitatively. Each of the 20 "bands," or groups of frequencies, were weighted according to how much of the articulation index was encompassed within the frequency groupings. This weighting is as follows: 1- band: 250–7000 Hz 100%; 2-band: 250–1500 Hz 40%, 1500–7000 60%; 3-band 250–800 Hz 20%, 800–1500 Hz 20%, 1500–7000 Hz 60%; 4-band: 250–800 Hz 20%, 800– 1500 Hz 20%, 1500–2500 Hz 20%, 2500–7000 Hz 40%. Each band was normalized to a maximum value of 1.00, weighted, and then averaged together to generate an ensemble response for each speech sound. The graphs were then 3-point smoothed.

These ensemble response plots were then examined for events relating to the features in the stimuli. This qualitative comparison will be made for each of the four pairs of speech sounds. Trying to perform quantitative comparisons between the histograms has proved overly difficult and meaningless, due to the large variation in ensemble responses between each condition. But,

because the main features like number, location and size of peaks are discernable visually, a qualitative analysis, and a qualitative analysis alone, was performed. Specifically, the number of initial "peaks" or acute elevations of auditory nerve fiber responses during the word-initial consonant will be recorded. It is expected that both /b/ and /p/ will elicit three, initial peaks. In addition, vowel-onset time will be examined, and will theoretically be how labial, alveolar, stop-consonants /b/ and /p/ are differentiated, if at all possible, even in the absence of reliable frequency information.

Results

Responses from the speech sounds /baeb/, /bab/, /paep/ and /pap/ when whispered or normally spoken were recorded from auditory nerve fibers from 18 chinchillas. Each auditory nerve fiber has a specific frequency to which it responds best (characteristic frequency) and a threshold, or minimum sound pressure level that must be reached before it produces an action potential. The distribution of characteristic frequencies and thresholds of the auditory nerve fibers sampled is shown in **Figure 1**. The charts in **Figure 2** show the same type of information, but specific to each speech sound. **Figure 1** shows a high density and focused distribution of auditory nerve fiber samples between 5,000 Hz and 10,000 Hz. The density of auditory nerve fiber samples is much less below 5,000 Hz. /Baeb/'s whispered presentation had the fewest successful action potential recordings at 66, and /baeb/'s normal presentation had the most at 82 action potential recordings (samples). **Figure 2** reflects a similar pattern as **Figure 1**, showing that high-frequency auditory nerve fibers were sampled more than mid- or low-frequency auditory nerve fibers.

The ensemble-average, time histograms showing auditory nerve fiber responses to /baeb/ at 60 decibels of peak-equivalent sound pressure level (dB pe SPL) both normal and whispered are shown in **Figure 3**. Each "initial peak," typically defined by a rapid increase of action potentials followed by a rapid decrease occurring during the consonant, is labeled as "P." **Figure 3a** shows three, initial peaks followed by a more uniform pattern of shorter peaks resulting from the first 120 ms of the presentation of normally-spoken /baeb/. The whispered presentation of /baeb/ at the same intensity shows the same pattern of three initial peaks, seen in **Figure 3b** between 5 ms and 25 ms. After 25 ms, the consonant has ended and the vowel begins shortly after, which presents as a relatively uniform "saw-tooth" pattern of spikes. The vowel extends past 120 ms, and therefore the second /b/ in /baeb/ is not visible in these figures. The part of the speech sound between the consonant ending and vowel's beginning, called the vowel-onset time, sometimes elicits unique patterns as well, and will be discussed later.

Almost the same presentation can be seen with /baeb/ at a higher intensity, at 70 dB pe SPL. There are three, initial peaks between 5 ms and 25 ms for the normally voiced and whispered versions of /baeb/, followed by a relatively homogenous "saw-tooth" pattern where the vowel begins, as seen in **Figure 4**. Although the third peak, "P3," in **Figure 4b** is not as high in magnitude as its preceding initial peaks and about the same height as the peaks during the vowel, it is in the same position as the third peak in /baeb/ normally voiced at 60 dB pe SPL in **Figure 3a**, and is therefore considered a peak.

At 90 dB pe SPL, however, there is only one peak between 5 ms and 25 ms, during the articulation of the consonant, which can be seen in **Figure 5**. The peaks in terms of shape and size are almost the same. There is, however, a small "peak," labeled as "P1" that appears between 2 ms and 5 ms of which the significance is not known. The number of peaks was still

consistent between normal and whispered speech and are more similar in shape and size than the peaks at 60 dB pe SPL or 70 dB pe SPL.

The ensemble average of responses for /pap/ revealed one, initial peak. The size, shape and location of the initial peak remains independent of speech production method. **Figure 6** shows similarly sized and located peaks at 60 dB pe SPL. In both the whispered and normally spoken versions of /pap/, a unique pattern, donated by "P*," can be seen after the consonant, and before the vowel, between 25- and 35 ms. This pattern is also apparent at higher intensities of the same speech sound. This post-consonant, pre-vowel pattern is not as visible in other figures, however. The significance of this pattern, and its appearance during the vowel-onset time, is not yet known.

Figure 7 shows the results at 70 dB pe SPL, which are almost the same as those at 60 dB pe SPL. /Pap/ at 90 dB pe SPL is similar to /baeb/ at 90 dB pe SPL in that it shows a pattern distinct from the 60- and 70 dB pe SPL histograms, as seen in **Figure 8**. Again, the 90 dB version shows one, large peak in the normally voiced presentation of the speech sound and a similar peak in the whispered presentation, but with a preceding, smaller peak, as seen at about 5 ms in **Figure 8b**.

Speech sounds /paep/ and /bab/ had a different number of peaks, or significantly different location and magnitude of peaks, between methods of speech production. An example is shown in **Figure 9** with /paep/. **Figure 9a**, the normally voiced version of /paep/, has one, initial peak reaching 0.25 average spikes per trial on a scale normalized to one. The whispered version, however, has one small peak reaching 0.25 spikes on average per trial on a normalized scale and one large peak reaching 0.6 spikes on average per trial. The initial pattern of peaks between the two methods of speech production for the same speech sound are not the same in number,

location or magnitude. This difference is also apparent at higher sound pressure levels, and also for the speech sound /bab/. An example of the responses recorded from /bab/ can be seen in **Figure 10.** /Bab/ normally voiced, as seen in **Figure 10a**, has three, initial peaks. **Figure 10b**, the whispered presentation of /bab/ also shows three peaks at about 0.5 and 0.3 spikes on average per trial on a normalized scale, although the peaks are at different locations and are different magnitudes between speech production methods. When stop consonant /b/ is paired with the vowel sound /ae/, seen in **Figure 4a**, or with vowel sound /a/, as seen in **Figure 10a**, the pattern of initial peaks does not change. The other presentations of /bab/ at 60 and 90 dB pe SPL yield results that also have peaks that are different sizes and in different numbers across methods of speech production.

Discussion

Of the four pairs of speech sounds, /baeb/ was the only speech sound entirely consistent with the hypothesis that whispered and normally voiced presentations of certain speech sound would elicit the same number of initial peaks in relatively the same places. While /pap/ also supported the hypothesis in the sense that it had a consistent number of peaks independent of method of speech production method, it only had one initial peak, instead of the three, as first published by Clarey et al. (2004). Speech sounds /paep/ and /bab/'s responses displayed clear initial peaks, but not the same presentation of initial peaks previously published or that was consistent with its whispered or normally voiced counterpart. While /bab/ at 70 dB pe SPL presented with three peaks in both normal and whispered speech, at least two of the three peaks are different magnitudes and occur at much different times than with the other form of speech,

and are therefore not considered the same three, initial peaks between speech production methods.

The results of /paep/ and /bab/ being inconsistent with other publications is most likely due to the uneven sampling of auditory nerve fibers with different characteristic frequencies and a low sample size. These possibilities, and others, will be explored further, later.

Uniform and consistent "saw-tooth" pattern peaks can be seen beginning at 45 ms in **Figure 6a** and are representative of the voicing of the vowel, when the speaker's vocal cords vibrate. In conjunction with the vocal-cord vibration, the articulation of the mouth, tongue and lips form a speech sound. This is consistent with the spectrogram and waveform plot also seen in **Figure 6a** which shows a distinct change at around 45 ms. These patterns are not as visible in the whispered versions of speech sounds, as seen in **Figure 6b**, however. As part of the definition of whispering, voicing does not take place, so a distinct change in timing response, frequency or amplitude is hard to detect, like in **Figure 6b**. Instead of vocal cord vibration, a whispered sound is made merely by exhaling air and articulating with the mouth, tongue and lips to form a unique speech sound. The lack of voicing makes it difficult to identify and therefore use vowel-onset time (VOT) as a reliable cue to differentiate between whispered speech sounds. Despite these obstacles, the number of initial peaks is, however, consistent across production method for /baeb/ and /pap/. Because both /p/ and /b/ result in three initial peaks, the number of peaks cannot be used alone to distinguish between them, as is possible with other consonants (Stevens & Wickesberg, 1999). The consistency of timing of auditory nerve fiber responses across frequencies and methods of speech production suggest timing information, in the absence of reliable frequency information, is a reliable cue and could be key in speech perception for these consonants. If a link could be found between timing patterns and vowel-onset time, perhaps the

two together could serve as a primary method of distinguishing between stop consonants, as theorized by Lisker in 1978.

The peaks in **Figure 3a**, and many other histograms, get smaller as time goes on because the auditory nerve fibers cannot respond as often, due to an increasingly depleted source of neurotransmitter from the inner- and outer-hair cells to the auditory nerve fibers, which the fibers need to produce an action potential. This also explains why the peaks after 20 ms, during the vowel of the speech sound, do not significantly increase in size even when the intensity of the speech sound is increased. For instance, the 60 dB pe SPL version of /baeb/ has initial peaks averaging about 0.37 spikes on average per trial on a normalized scale. With increased intensity at 70 dB pe SPL, the peaks during the vowel average about 0.13 spikes on average per trial on a normalized scale and decrease over time. Finally, at 90 dB, the peaks during the vowel are again around 0.15 and decrease over time. So, even with a 30 dB increase, the spikes on average per trial do not increase significantly indicating they are most likely unable to due to the hair cells limit to how much neurotransmitter it can release per unit time. If the hair cells could produce and release more neurotransmitter than it does at 60 dB pe SPL, the peaks would increase in magnitude as SPL increased, as the initial peaks did between 60- and 90 dB pe SPL, which they do not.

Although whispered /baeb/ and normally voiced /baeb/ both have three peaks between 5 ms and 20 ms, they are in slightly different places, are different magnitudes and are slightly different shapes. This could be because the whispered and normally voiced presentations come from different recordings, among other things. Each recording has its own sound pressure level, vowel-onset time, emphasis, articulation, and so forth. Seeing differences in the presentation of the peaks is likely due to these variables (Tartter, 1989). Since whispered speech is not naturally

at 60-, 70- or 90 dB pe SPL, the recordings of the speech sounds were modified to have peak equivalents to these values, as were the normally voiced recordings. When normally spoken, the voicing of the vowel tends to be the loudest part of the speech sound and therefore determines the peak-equivalent SPL. In whispered speech, however, the consonant tends to be the loudest component and in that case determines the peak-equivalent SPL. This discrepancy makes it difficult to make each pair of speech sounds, normally voiced and whispered, of equal intensity.

The normally voiced version of the speech sound might differ from the whispered in that the consonants are more articulated in the whispered version, causing a slightly different pattern of peaks at the beginning of the speech sound. In addition, the whispered and normally voiced versions have different formants and formant transitions, or prominent frequencies, and the non-uniform distribution of auditory nerve fibers with characteristic frequencies lead one histogram to be different from another. This implies that looking at more than just the number of initial peaks found from averaging auditory nerve fibers responses may be necessary since so many variables affect the presentation and reliability of responses to speech sounds with different speech production styles. In addition to the consonant, the characteristics of the vowel in the speech sound might differ between production methods and elicit different responses from the auditory nerve fibers. The formants, which come from the vowel, as opposed to the formant transitions, coming from the consonant, are different in whispered speech (Tartter, 1989). Since few studies have examined whispered speech and how to best analyze it, perhaps the importance of prominent frequencies being different in whispered speech has not yet been understood and the method of analysis should be different.

The distribution of characteristic frequency and threshold in **Figure 1** reveals the unrepresentative sampling of auditory nerve fibers from the chinchillas, as the graph is most

dense from 5,000 Hz to 10,000 Hz. The frequencies most important to understanding speech are between 250 Hz and 7,000 Hz, which capture the majority of the fibers sampled, however not uniformly (French & Steinberg, 1947). This uneven distribution of auditory nerve fiber characteristic frequency is most likely due to how the fibers are organized when exiting the cochlea. The auditory nerve fibers twist around each other with the high-frequency fibers more superficial and the lower-frequency auditory nerve fibers more deep. This means when the auditory nerve is sampled, it is much easier to access the high-frequency fibers with the electrode than the buried, low-frequency fibers. While high-frequency fibers are useful, especially for whispered speech, which tends to have predominantly higher frequencies in it, each of the 20 groups of frequencies is said to contribute equally in the perception of speech (French & Steinberg, 1947). Having an undersampling of lower frequencies might therefore lead to unequal, or absent, representation of the prominent frequencies in the speech sound. To try and compensate for this, each of the 20 bands was normalized before they were averaged and plotted in the ensemble average peri-stimulus time histogram. Some bands, however, only had 1 or 2 auditory nerve fibers, while others, typically of higher frequency, sometimes had 12 bands. Since each band represents 1/20 of the overall frequency range, or 5 percent, if a band only had data from one auditory nerve fiber, it represented a disproportionate amount of all frequencies. On the other hand, if a band had 10 auditory nerve fibers sampled within it, each nerve fiber would then represent 1/200 of the frequency range, making the normalized value for that band potentially much more accurate. The Articulation Index, which states that each band represents an equal portion of frequencies between 250 Hz – 7000 Hz, was developed for normally voiced speech. The Articulation Index may not be appropriate for whispered speech, since whispered speech contains different frequencies than normally voiced speech. Whispered speech, for instance,

tends to involve much higher frequencies than does normally voiced speech, including frequencies above the Articulation Index's upper limit of 7000 Hz. This can be seen in the spectrograms in any of the speech sounds, like with /baeb/. The whispered version, in **Figure 3b**, has frequencies well above 7000 Hz, whereas the normally voiced version has almost no frequency information above 6000 Hz, as seen in **Figure 3a**. So, perhaps using a separate "index" for whispered speech that more accurately divides each band of frequencies would allow for more accurate normalizing.

Figure 2 shows the distributions of auditory nerve fibers sampled for each speech sound. While the density and distribution is similar to the overall distribution shown in **Figure 1**, not every speech sound, or band within it, has the same number of fibers sampled. Normally voiced /baeb/, which yielded results consistent with the hypothesis and previously published material, had the highest number of auditory nerve fibers sampled in response to it at 82 samples. The whispered version of /baeb/ had the least, at 66 samples. While the histogram for /baeb/ whispered is still clear, it is not as clear as its normally voiced counterpart implying that clarity and perhaps reliability of timing information increases with sample size.

The responses elicited by /pap/ also supported part of the hypothesis in that the number, size and shape of initial peaks were consistent across methods of speech production, as seen in **Figure 6**. There is, however, only one peak instead of three, as published by Clarey et al. (2004). Even though /p/ and /b/ both were expected to produce three initial peaks, they were expected to be distinguishable by vowel-onset time, or the time between the consonant is finished and the vowel begins during speech production. In normally spoken speech, it is often clear where the voice- or vowel-onset time begins because of a notable change in peak pattern, as noted in

Figure 6a by the "VOT" arrow. In whispered speech, however, it is much more difficult to find in the timing pattern, although sometimes still recognizable in the speech sound's spectrogram.

Not only do the auditory nerve fiber responses exhibit consistent "saw-tooth" patterns, but it can be seen in the corresponding wave-form plot and spectrogram that amplitude and frequency changes, indicating the beginning of voicing (Lisker 1978 & Kiang et al., 1965). During whispered speech, however, there is no voicing. This makes differentiating between the consonant, a pause, and vowel-onset time difficult. Sometimes the consonant, vowel and time between them each have unique patterns, but this seems to be rare. An instance of each having a unique, distinguishable timing pattern, again, is **Figure 6**. The consonant results in one, prominent, initial peak, followed by a unique, 3-peaked pattern labeled as "P*" and finally the saw-toothed pattern characteristic of the vowel. Because the exact vowel-onset time time location is not reliable or easily identifiable in whispered speech, using it to differentiate between /b/ and /p/ is not possible with the responses recorded. In addition, the whispered versions of /p/ and /b/ seem to have significantly different vowel-onset times than their normally-voiced counterparts, as seen in their spectrograms. The vowel sound /ae/ in /paep/, for instance, seen in Figure 9, seems to start around 85 ms in the normally voiced version, but at about 2 ms in the whispered version. The vowel-onset time can be discerned not only by the sudden presence of frequencies (colored red and yellow) seen in their respective spectrogram, but also by the rapid increase in amplitude, seen in the waveform representation of the speech sound. The extended vowel-onset time time can likely be attributed to the task of vibrating the vocal cords, which is not done in the whispered version of the speech sound. Since /p/ and /b/ are stop consonants, the speaker must stop the air to finish the consonant, and then consistently vibrate the vocal cords to produce the voicing of the vowel, in this case, /ae/. While initiating the vibration of vocal cords

is a quick process, it is not as quick as merely articulating the vowel by shaping the mouth, tongue and lips while exhaling, to produce the whispered version of /ae/. This discrepancy in time could account for the VOT difference between voiced and unvoiced (whispered) vowel-onset times.

The vowel-onset time differences across speech production methods and lack of voicing in whispered speech makes differentiating between vowels difficult, as both vowel sounds /ae/ and /a/, when paired with a consonant, seemed to elicit the same initial peaks and timing pattern. As seen in **Figure 4a** and **Figure 10a**, both vowels evoke three, initial peaks regardless of which vowel proceeds the consonant. The timing pattern for the vowel also seems indistinguishable, as both present as a repetitive "saw-tooth" pattern after the consonant. So, neither the timing nor the pattern of vowel-onset time seem to be adequate in differentiating between /p/ and /b/ consonants. Prominent frequencies, or formants, also seem to be dependant on vowel sounds and not on the consonant, which is consistent with Peterson and Barney's description of formants (1952). Formant transitions, or small change in frequencies preceding the formant's consistent frequency, however, are meant to allow differentiation between consonants. These formant transitions, however are not visible enough in the spectrograms from our data, and therefore cannot be analyzed. Different formants can be seen, however, between vowel sounds /ae/ and /a/, in some instances of whispered speech, as seen in **Figure 7b** and **9b**. The formants resulting from vowel sound /a/ are not the same in **Figures 7b** and **10b**, even though they both represent speech sounds with vowel sound /a/. This inconsistency among formants despite the presence of the same vowel sounds is unexplained by Liberman, who claimed vowel sounds would have consistent frequencies, regardless of its preceding consonant (1967). While Liberman's claim is often true for normally spoken speech, as seen in the spectrograms from speech sounds in this

study, the same system does not hold true for whispered speech. It is, however, unlikely that there is a separate system involving formant recognition for understanding whispered speech, as no previous study has been found to suggest infants learn whispered speech separately from normally spoken speech, suggesting one, common mechanism is used to perceive these, and any other kind of speech, like sung or shouted speech. The common mechanism could be rooted in timing patterns. While frequencies might contribute to speech perception, they are unlikely to play a large role, as evidenced by these data.

Other instances and intensities of speech sound recordings did not necessarily elicit responses consistent with the hypothesis, often because the number of initial peaks was not consistent across speech-production method. Aside from the initial "peak," that appears in whispered /baeb/ between 2 and 7 ms whose height is shorter than any other initial peak in any of the histograms, the normal and whispered versions of /baeb/ elicit one, similarly shaped initial peak at about 0.55 spikes on average per trial on a normalized scale between 5 ms and 15 ms as seen in **Figure 5**. The small "peak" that is easily observable at 60- and 70 dB pe SPL and appears at 90 dB pe SPL in **Figure 5b**, marked by "P1," is most likely an artifact in the sound recording. The recording at 90 dB pe SPL is 20-30 dB higher than the others. Since something that is 30 dB pe SPL louder is perceived as 6 times as loud because of Stevens' Power Law, an artifact causing 0.1 spikes on average per trial can get intensified to .25 spikes on average per trial on a normalized scale, as seen between **Figure 3a** and **Figure 5a**. Since this "peak" is most likely an artifact of the sound recording, and not a significant response to the consonant in the speech sound it could be disregarded.

One possible explanation of /baeb/ at 90 dB pe SPL eliciting different results than at 60- or 70 dB pe SPL is because many of the auditory nerve fibers are saturated, and above their

threshold at 90 dB pe SPL. **Figure 1** illustrates the highest threshold any of the auditory nerve fibers reach is about 42 dB pe SPL. The dynamic response range, or the range of intensity to which each fiber responds, is 30 dB SPL (Kiang et al., 1965). Although the sound pressure level published by Kiang (1965) is on a slightly different scale than the peak equivalent sound pressure level scale, it implies that the range to which the auditory nerve fibers sampled should be at most 72 dB pe SPL. It is therefore understandable that at 90 dB pe SPL, results are not as reliable, as seen in **Figure 5** for 90 dB pe SPL /baeb/, which has results inconsistent with the same speech sounds at 60- and 70 dB pe SPL.

Other characteristics, however, like the location of the peaks, size and width could be used to differentiate between the two alveolar stop consonants. Because the expected, consistant three peaks were not seen in the results for /pap/ or /paep/, exploring this method is not possible. The reason why there were not three peaks or consistency across methods of speech production is most likely due to sampling that was not uniform across frequencies, as seen in **Figure 1**. The low-frequency, low sound-pressure-level threshold portion of the scatter plot is expected to be empty, as the hearing threshold of the chinchilla and human does not occur until higher frequencies and thresholds for sound pressure levels.

While each intensity level for each speed sound contained data from mostly the same auditory nerve fibers, data from each pair of speech sounds, normally spoken and whispered, were not necessarily from the same set of auditory nerve fibers. Recording data from the same auditory nerve fibers with both speech production methods could have eliminated some variables and perhaps increased the reliability of the number of initial peaks between speech production methods. Other variables, such as speaker, most likely do not have a significant effect on results as noted by Stevens and Wickesberg (1999). Stevens and Wickesberg discussed the timing

patterns found in normally spoken and whispered /t/ and /d/, and this study examined /p/ and /b/ (1999). Future studies can examine other categories of consonants, such as nasals (/m/ and /n/) or fricatives (/s/ and /z/). Furthermore, vowels can be examined for timing patterns, or other forms of speech other than whispered, such as sung or shouted speech. The terminal consonant can also be examined for timing patterns, as in the second /b/ in the speech sound /baeb/. While neurotransmitter depletion and the preceding vowel might affect its pattern in a way that the word-initial consonant is not, it might still provide useful information. Preliminary analysis of word-final consonants with the current data has shown three peaks at the onset of the second consonant similar to the initial peaks for the word-initial /b/ for /bab/. This analysis could be done for all the intensity levels for each speech sound to see if similar results arise. Previous studies have shown cochlear nucleus neurons to have similar temporal patterns as those from the auditory nerve, and is something that could be analyzed again in future studies concerning temporal patterns like this one (Clarey et al., 2004).

Until recently, it was generally accepted that frequency was the main cue in understanding speech. Studies like this one or others can hopefully shed light on the General Auditory Theory, which posits there must be cues other than frequency that allow us to perceive speech. Studies like this that show cues, like timing, as important in understanding speech can help develop the General Auditory Theory and hopefully lead to advancements to assistive hearing technology and speech therapy.

Despite all the obstacles to perceiving speech like varying frequencies, background noise and conversations and loud instruments, speech is all somehow still perceivable. How the auditory system uses timing patterns, even with the help of frequency cues and voice-onset time is still far from understood. Understanding further what is thought to be the key components of

speech sounds, frequency and timing, will hopefully lead to answers as to how the auditory system turns vibrating air into meaningful language.

References

Clarey, J.C., Paolini, A.G., Grayden, D.B., Burkitt, A.N., Clark, G.M. 2004. Ventral cochlear nucleus coding of voice onset time in naturally spoken syllables. *Hearing Research*, 190, 37-59.

Corti, A. 1851. Recherches sur l'organe de Corti de l'ouïe des mammifères. *Zwiss Zool,* 3, 1-106.

Dannenbring, G. L. 1980. Perceptual Discrimination of Whispered Phoneme Pairs. *Perceptual & Motor Skills,* 51, 979-985.

Diehl, Randy L., Lotto, A.J., Holt, L.L. 2004. Speech Perception. *Annual Review of Psychology*, 55, 149-79.

Delattre, Pierre C., Liberman, A.M., Cooper, Fanklin S. 1955. *Journal of The Acoustical Society of America,* 27, 769-73.

Denes, P. B., Elliot P.N. 1993. *The Speech Chain. The Physics and Biology of Spoken Language*. New York, NY: W.H. Freeman and Company.

French, N.R., Steinberg, J.C. 1947. Factors governing the intelligibility of speech sounds. *The Journal of The Acoustical Society of America,* 19, 90-119.

Kiang NYS, Watanabe T., Thomas C., Clark L.F. 1965. Discharge Patterns of Single Fibers in the Cat's Auditory Nerve. Cambridge, MA: MIT Press.

Kuhl, Patricia K., Miller, JD. 1975. Speech Perception by the Chinchilla: Voiced-Voiceless Distinction in Alveolar Plosive Consonants. *Science*, 190, 69-72.

Liberman, A.M. 1957. Some results of research on speech perception. *The Journal of The Acoustical Society of America,* 29, 117–23.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P. 1967. Perception of the Speech Code. *Psychological Review,* 74, 431-461.

Liberman, A.M., DeLattre, P.D., Cooper, F.S. 1952. The role of selected stimulus variables in the perception of the unvoiced stop consonants. *American Journal of Psychology*, 65, 497-516.

Liberman, A.M., Mattingly, I.G., 1985. The motor theory of speech perception revised. *Cognition*, 21, 1-36.

Lisker, Leigh. 1978. In qualified defense of VOT. *Language and Speech*, 21, 375-383.

Loebach, Jeremy L., Wickesberg, R.E. 2006. The representation of noise vocoded speech in the auditory nerve of the chinchilla: Physiological correlates of the perception of spectrally reduced speech. *Hearing Research*, 213, 130-44.

Peterson, Gordon E., Barney, H. 1952. Control Methods Used in a Study of the Vowels. *The Journal of The Acoustical Society of America*, 24,175-84.

Repp, B.H., Lin, H.B. 1989. Acoustic properties and perception of stop consonant release Transients. *The Journal of The Acoustical Society of America,* 85, 379-396.

Schwartz, M.F., 1970. Power spectral density measurements of oral and whispered speech. *Journal of Speech and Hearing Research,* 13, 445-446.

Simmons, F.B. 1966. Electrical Stimulation of the Auditory Nerve in Man. *Archives of Otolaryngology,* 84, 2-54.

Shannon, R.V., Zeng, F., Kamath, V., Wygonski, J., Ekelid, M.,1995. Speech recognition with primarily temporal cues. *Science,* 270, 303-4.

Stevens, H., Wickesberg, R. E. 1999. Ensemble responses of the auditory nerve to normal and whispered stop consonants. *Hearing Research*, 131, 47-62.

Stevens, K.N., Blumstein, S.E. 1981. The search for invariant correlates of phonetic features, Perspectives on the study of speech, eds P.D. Eimas and J.L. Miller, Erlbaum; Hillsdale, N.J.

Tartter, Vivien. 1989. What's in a Whisper? *Journal of the Acoustical Society of America*, 86, 1678-83.

Figure 1. Scatter plot of the distribution of the characteristic frequency and threshold of the 174 sampled auditory nerve fibers from the 18 chinchillas used in this study.

Figure 2. Scatter plots of the distributions of the characteristic frequencies and thresholds of sampled auditory nerve fibers specific to each speech sound.

Figure 3. Ensemble responses and spectrograms for the first 120 ms of normally spoken (3a) and whispered (3b) /baeb/ presented at 60 dB pe SPL. The normalized, average number of action potentials are plotted on the y-axis and time is plotted on the x-axis. Each peak during the consonant is labeled with an arrow and numbered. Each spectrogram shows a visual representation of the first 120 ms of the speech sound, red being the highest intensity frequencies and white the lowest. A time waveform is plotted below the spectrogram.

Figure 4. Ensemble responses and spectrograms for the first 120 ms of normally spoken (4a) and whispered (4b) /baeb/ presented at 70 dB pe SPL. The normalized, average number of action potentials are plotted on the y-axis and time is plotted on the x-axis. Each peak during the consonant is labeled with an arrow and numbered. Each spectrogram shows a visual representation of the first 120 ms of the speech sound, red being the highest intensity frequencies and white the lowest. A time waveform is plotted below the spectrogram.

Figure 5. Ensemble responses and spectrograms for the first 120 ms of normally spoken (5a) and whispered (5b) /baeb/ presented at 90 dB pe SPL. The normalized, average number of action potentials are plotted on the y-axis and time is plotted on the x-axis. Each peak during the consonant is labeled with an arrow and numbered. Each spectrogram shows a visual representation of the first 120 ms of the speech sound, red being the highest intensity frequencies and white the lowest. A time waveform is plotted below the spectrogram.

Figure 6. Ensemble responses and spectrograms for the first 120 ms of normally spoken (6a) and whispered (6b) /pap/ presented at 60 dB pe SPL. The normalized, average number of action potentials are plotted on the y-axis and time is plotted on the x-axis. Each peak during the consonant is labeled with an arrow and numbered. Each spectrogram shows a visual representation of the first 120 ms of the speech sound, red being the highest intensity frequencies and white the lowest. A time waveform is plotted below the spectrogram.

Figure 7. Ensemble responses and spectrograms for the first 120 ms of normally spoken (7a) and whispered (7b) /pap/ presented at 70 dB pe SPL. The normalized, average number of action potentials are plotted on the y-axis and time is plotted on the x-axis. Each peak during the consonant is labeled with an arrow and numbered. Each spectrogram shows a visual representation of the first 120 ms of the speech sound, red being the highest intensity frequencies and white the lowest. A time waveform is plotted below the spectrogram.

Figure 8. Ensemble responses and spectrograms for the first 120 ms of normally spoken (8a) and whispered (8b) /pap/ presented at 90 dB pe SPL. The normalized, average number of action potentials are plotted on the y-axis and time is plotted on the x-axis. Each peak during the consonant is labeled with an arrow and numbered. Each spectrogram shows a visual representation of the first 120 ms of the speech sound, red being the highest intensity frequencies and white the lowest. A time waveform is plotted below the spectrogram.

Figure 9. Ensemble responses and spectrograms for the first 120 ms of normally spoken (9a) and whispered (9b) /paep/ presented at 60 dB pe SPL. The normalized, average number of action potentials are plotted on the y-axis and time is plotted on the x-axis. Each peak during the consonant is labeled with an arrow and numbered. Each spectrogram shows a visual

representation of the first 120 ms of the speech sound, red being the highest intensity frequencies and white the lowest. A time waveform is plotted below the spectrogram.

Figure 10. Ensemble responses and spectrograms for the first 120 ms of normally spoken (10a) and whispered (10b) /bab/ presented at 70 dB pe SPL. The normalized, average number of action potentials are plotted on the y-axis and time is plotted on the x-axis. Each peak during the consonant is labeled with an arrow and numbered. Each spectrogram shows a visual representation of the first 120 ms of the speech sound, red being the highest intensity frequencies and white the lowest. A time waveform is plotted below the spectrogram.

Figure 1

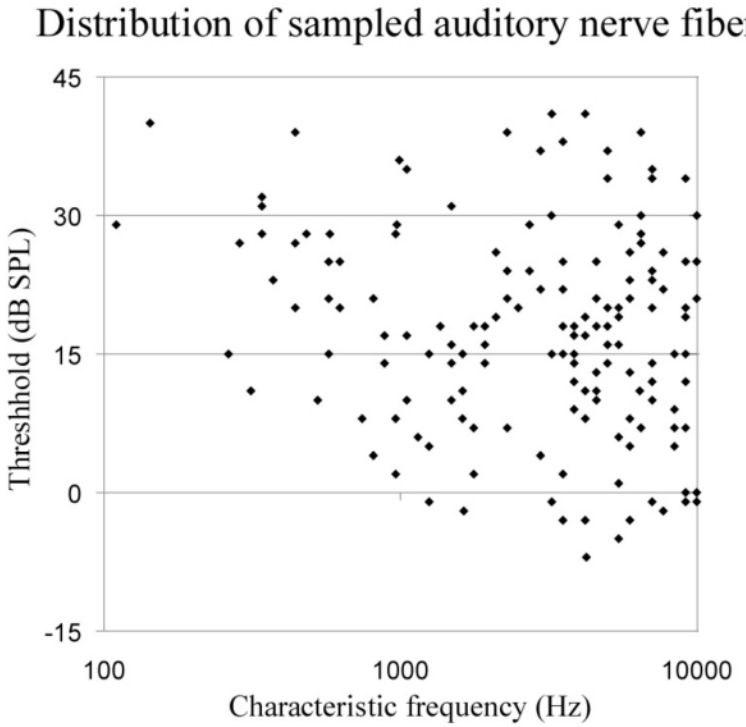

Figure 2

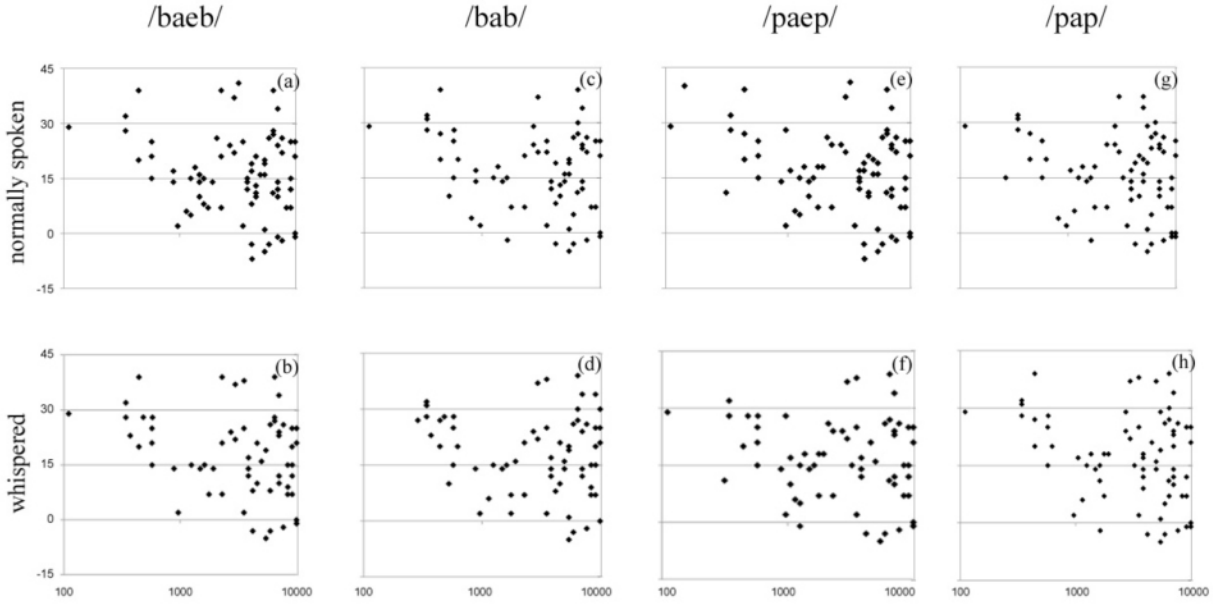

Figure 3a

/baeb/ normally voiced 60 dB pe SPL

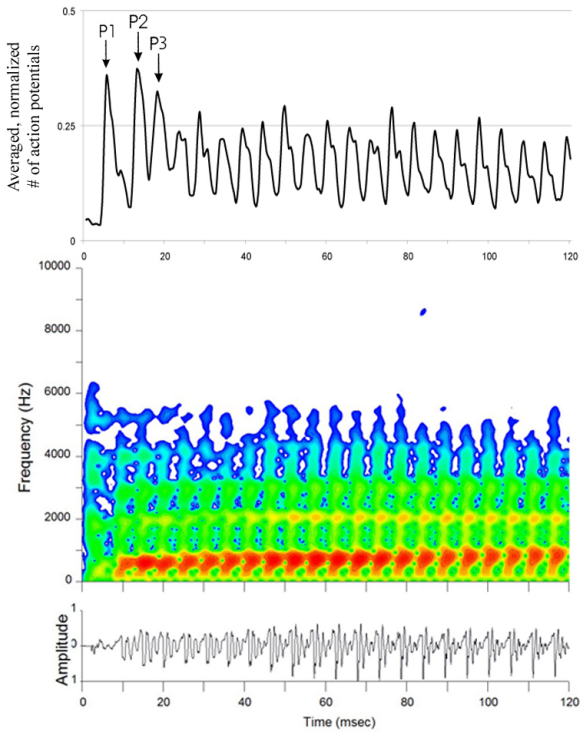

Figure 4a

/baeb/ normally voiced 70 dB pe SPL

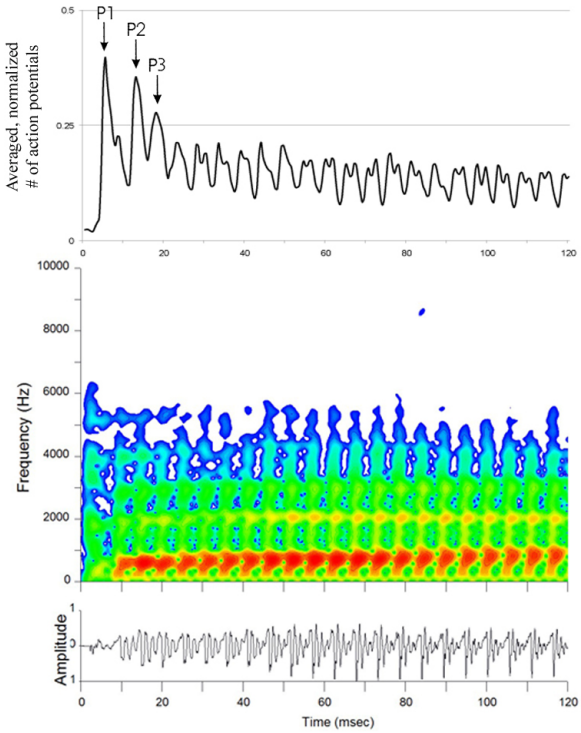

Figure 3b

/baeb/ whispered 60 dB pe SPL

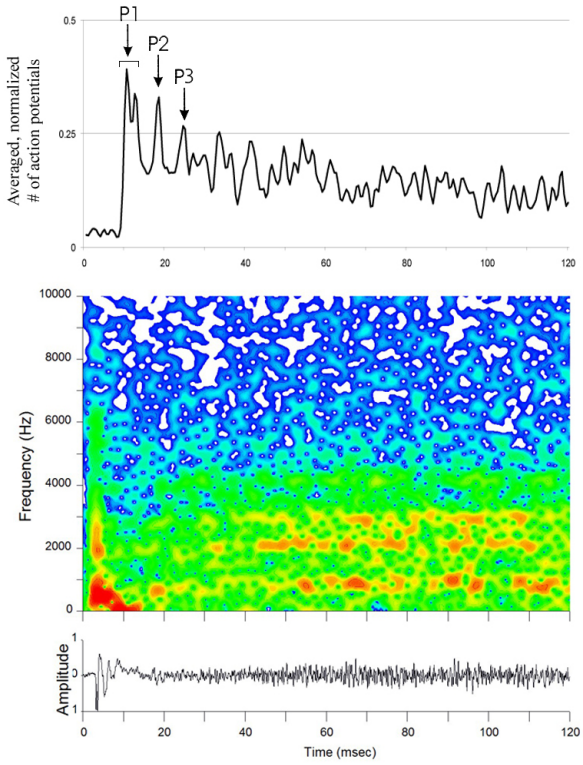

Figure 4b

/baeb/ whispered 70 dB pe SPL

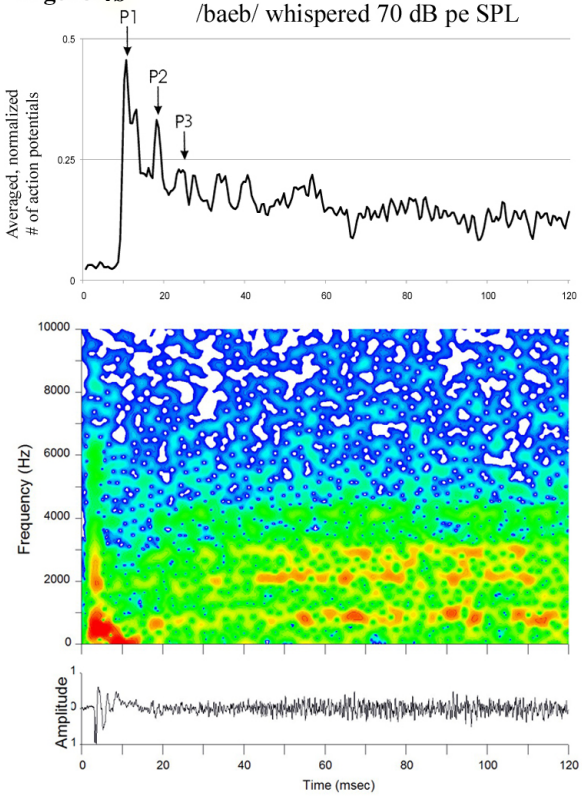

Figure 5a

/baeb/ normally voiced 90 dB pe SPL

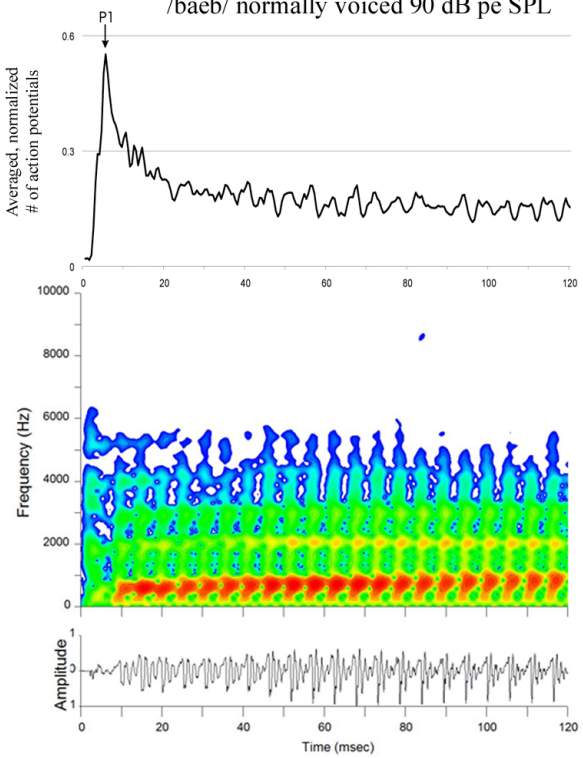

Figure 6a

/pap/ normally voiced 60 dB pe SPL

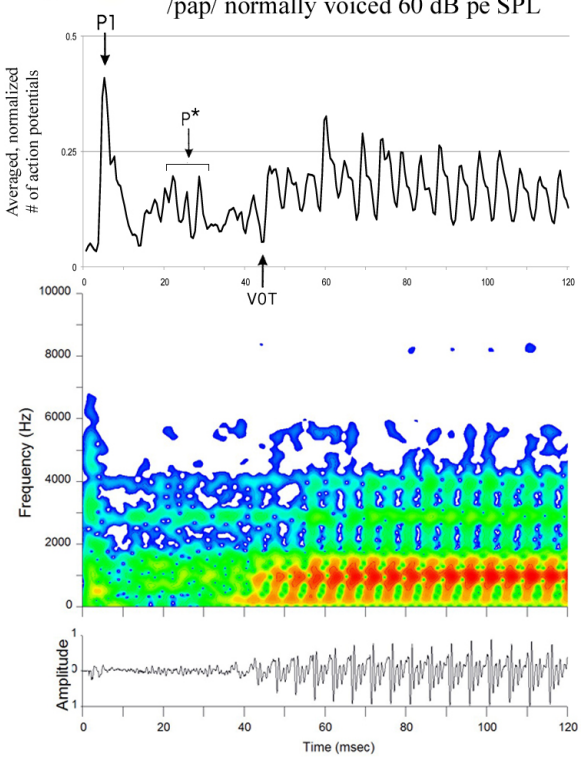

Figure 5b

/baeb/ whispered 90 dB pe SPL

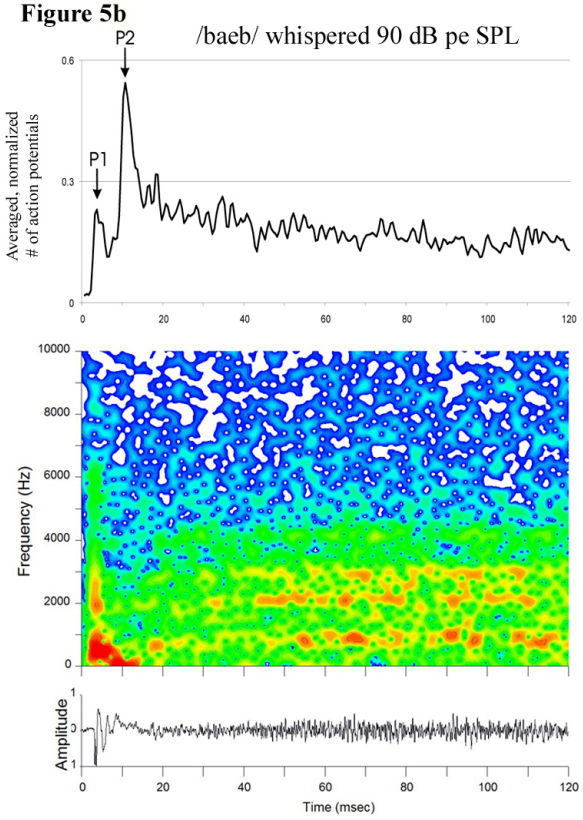

Figure 6b

/pap/ whispered 60 dB pe SPL

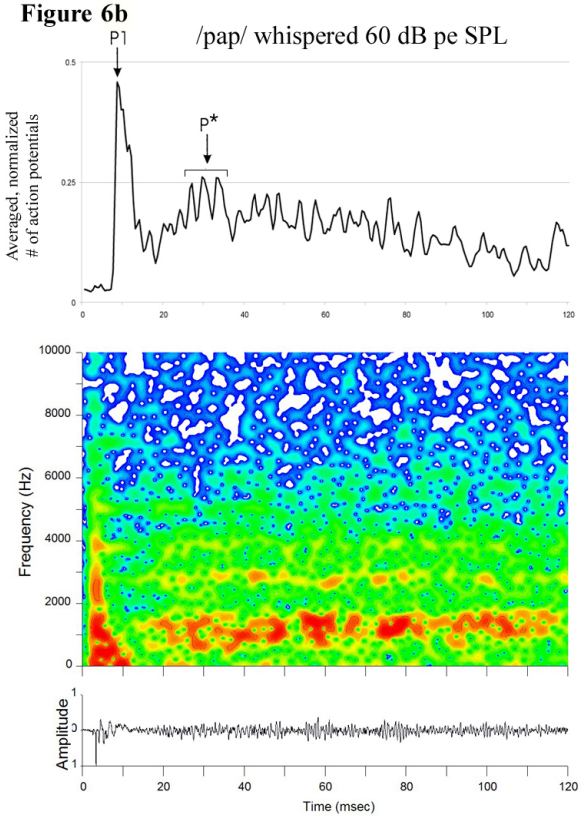

Figure 7a

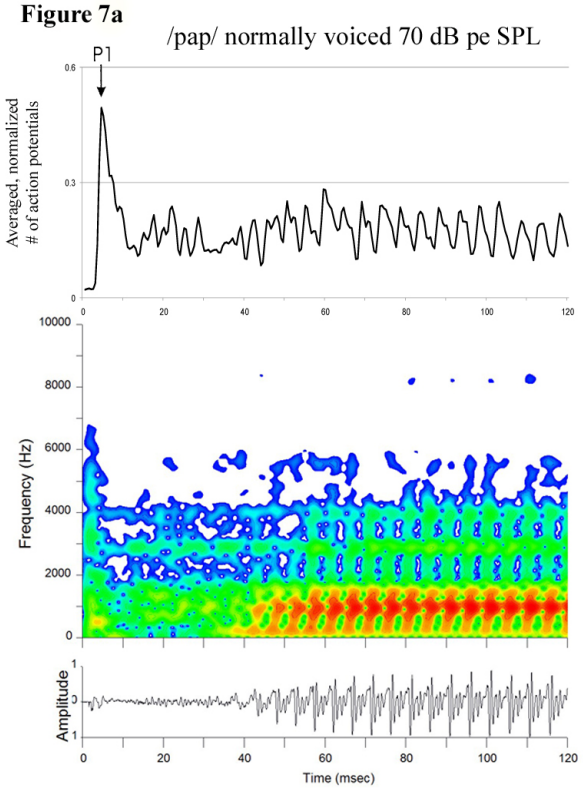

Figure 8a

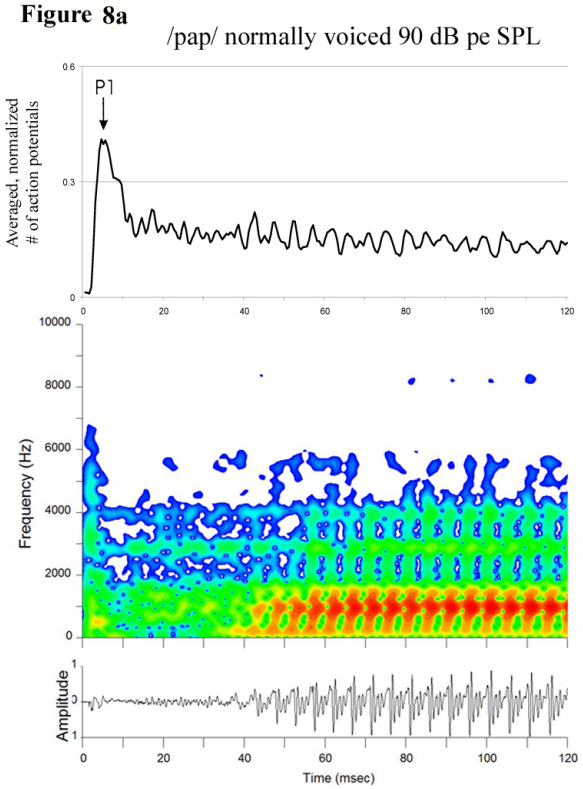

Figure 7b

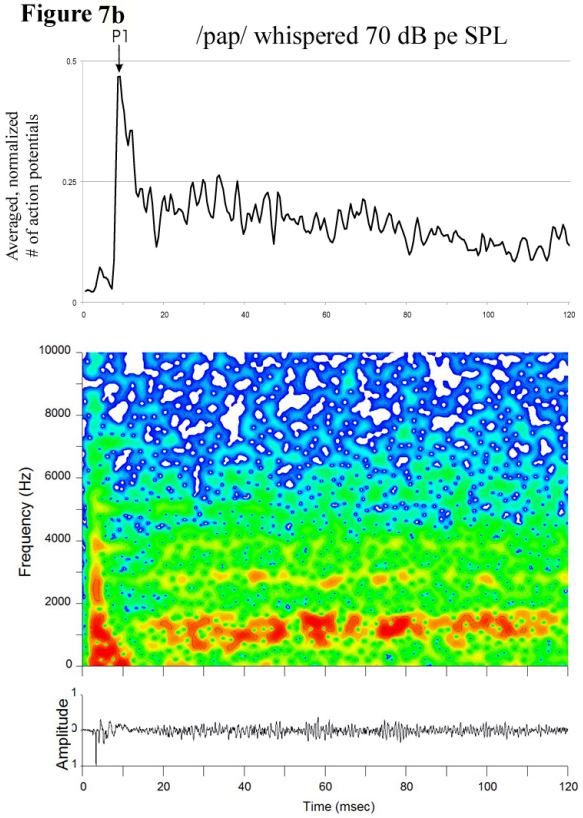

Figure 8b

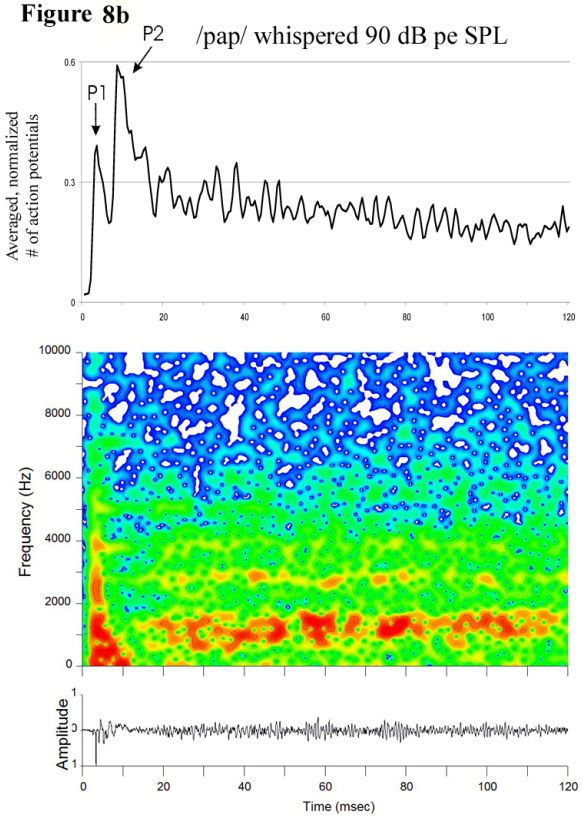

Figure 9a

/paep/ normally voiced 60 dB pe SPL

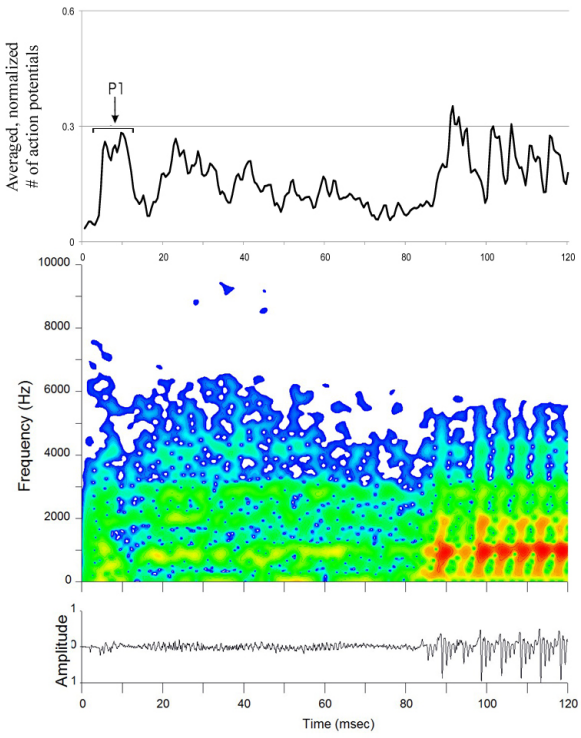

Figure 10a

/bab/ normally voiced 70 dB pe SPL

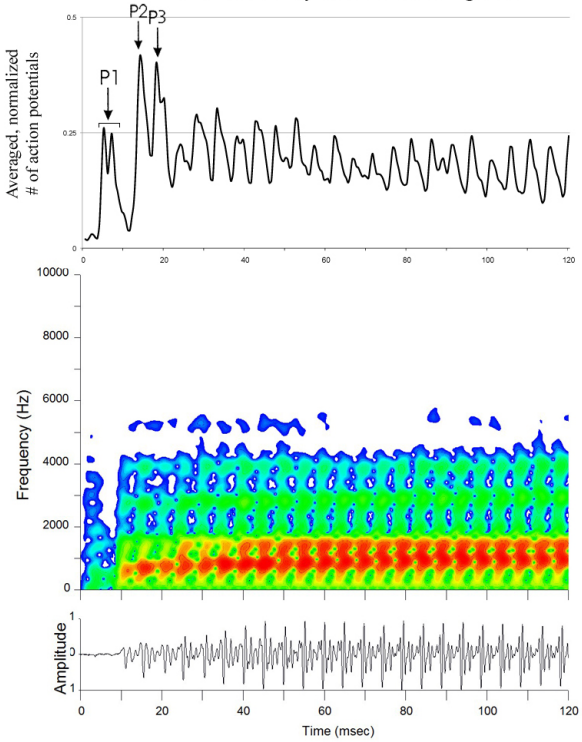

Figure 9b

/paep/ whispered 60 dB pe SPL

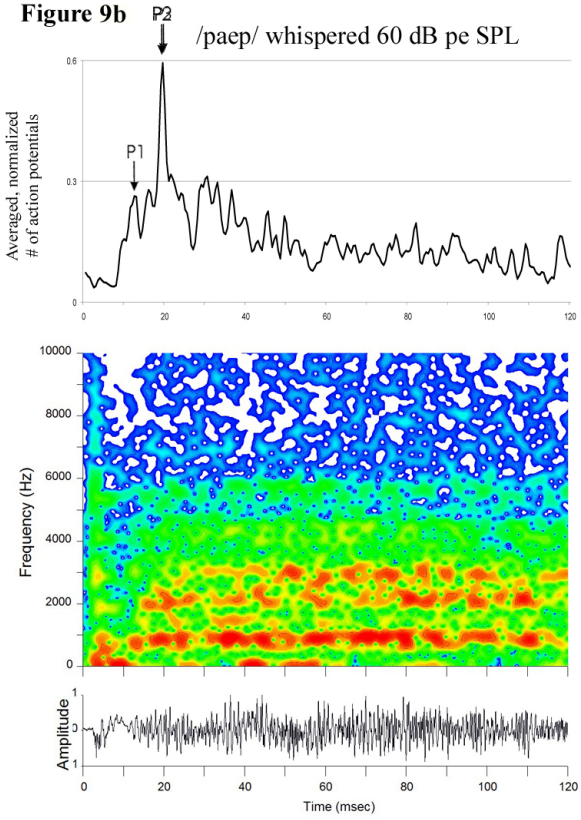

Figure 10b

/bab/ whispered 70 dB pe SPL